Paraphrasing Revisited with Neural Machine Translation

Paraphrasing Revisited with Neural Machine Translation

Jonathan Mallinson, Rico Sennrich and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
J.Mallinson@ed.ac.uk, {rsennric,mlap}@inf.ed.ac.uk

Abstract

Recognizing and generating paraphrases is an important component in many natural language processing applications. A well-established technique for automatically extracting paraphrases leverages bilingual corpora to find meaning-equivalent phrases in a single language by "pivoting" over a shared translation in another language. In this paper we revisit bilingual pivoting in the context of neural machine translation and present a paraphrasing model based purely on neural networks. Our model represents paraphrases in a continuous space, estimates the degree of semantic relatedness between text segments of arbitrary length, or generates candidate paraphrases for any source input. Experimental results across tasks and datasets show that neural paraphrases outperform those obtained with conventional phrase-based pivoting approaches.

1 Introduction

Paraphrasing can be broadly described as the task of using an alternative surface form to express the same semantic content (Madnani and Dorr, 2010). Much of the appeal of paraphrasing stems from its potential application to a wider range of NLP problems. Examples include query and pattern expansion (Riezler et al., 2007), summarization (Barzilay, 2003), question answering (Lin and Pantel, 2001), semantic parsing (Berant and Liang, 2014), semantic role labeling (Woodsend and Lapata, 2014), and machine translation (Callison-Burch et al., 2006).

Most of the recent literature has focused on the automatic extraction of paraphrases from various different types of corpora consisting of parallel, non-parallel, and comparable texts. One of the most successful proposals uses bilingual parallel corpora to induce paraphrases based on techniques from phrase-based statistical machine translation (SMT, Koehn et al. (2003)). The intuition behind Bannard and Callison-Burch's (2005) bilingual pivoting method is that two English strings e_1 and e_2 that translate to the same foreign string f can be assumed to have the same meaning. The method then pivots over f to extract $\langle e_1, e_2 \rangle$ as a pair of paraphrases. Drawing inspiration from syntax-based SMT, several subsequent efforts (Callison-Burch, 2008; Ganitkevitch et al., 2011) extended this technique to syntactic paraphrases leading to the creation of PPDB (Ganitkevitch et al., 2013; Ganitkevitch and Callison-Burch, 2014), a large-scale paraphrase database containing over a billion of paraphrase pairs in 23 different languages.

In this paper we revisit the bilingual pivoting approach from the perspective of neural machine translation, a new approach to machine translation based purely on neural networks (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Sutskever et al., 2014; Luong et al., 2015). At its core, NMT uses a deep neural network trained end-to-end to maximize the conditional probability of a correct translation given a source sentence, using a bilingual corpus. NMT models have obtained state-of-the art performance for several language pairs (Jean et al., 2015b; Luong et al., 2015), using only parallel data for training, and minimal linguistic information. In this paper we show how the bilingual pivoting method can be ported to NMT and argue that it offers at least three advantages over conventional methods. Firstly, our neural paraphrasing model learns continuous space representations for phrases and sentences (aka *embeddings*) that can be usefully incorporated in downstream tasks such as recognizing textual similarity and entailment. Secondly, the proposed model is able to either score a pair of paraphrase candidates (of arbitrary length) and generate target paraphrases for a given source input. Due to the architecture of NMT, generation takes advantage of wider context compared to phrase-based approaches: target paraphrases are predicted based on the meaning of the source input and all previously generated target words.

In the remainder of the paper, we introduce our

paraphrase model and experimentally compare it to the phrase-based pivoting approach. We evaluate the model's paraphrasing capability both *intrinsically* in a paraphrase detection task (i.e., decide the degree of semantic similarity between two sentences) and *extrinsically* in a generation task. Across tasks and datasets our results show that neural paraphrases yield superior performance when assessed automatically and by humans.

2 Related Work

The literature on paraphrasing is vast with methods varying according to the type of paraphrase being induced (lexical or structural), the type of data used (e.g., monolingual or parallel corpus), the underlying representation (surface form or syntax trees), and the acquisition method itself. For an overview of these issues we refer the interested reader to Madnani and Dorr (2010). We focus on bilingual pivoting methods and aspects of neural machine translation pertaining to our model. We also discuss related work on paraphrastic embeddings.

Bilingual Pivoting Paraphrase extraction using bilingual parallel corpora was proposed by Bannard and Callison-Burch (2005). Their method first extracts a bilingual phrase table and then obtains English paraphrases by pivoting through foreign language phrases. Paraphrases for a given phrase are ranked using a paraphrase probability defined in terms of the translation model probabilities $P(f|e)$ and $P(e|f)$ where f and e are the foreign and English strings, respectively.

Motivated by the wish to model sentential paraphrases, follow-up work focused on syntax-driven techniques again within the bilingual pivoting framework. Extensions include representing paraphrases via rules obtained from a synchronous context free grammar (Ganitkevitch et al., 2011; Madnani et al., 2007) as well as labeling paraphrases with linguistic annotations such as CCG categories (Callison-Burch, 2008) and part-of-speech tags (Zhao et al., 2008).

In contrast, our model is syntax-agnostic, paraphrases are represented on the surface level without knowledge of any underlying grammar. We capture paraphrases at varying levels of granularity, words, phrases or sentences without having to *explicitly* create a phrase table.

Neural Machine Translation There has been a surge of interest recently in repurposing sequence transduction neural network models for machine translation (Sutskever et al., 2014). Central to this approach is an encoder-decoder architecture implemented by recurrent neural networks. The encoder reads the source sequence into a list of continuous-space representations from which the decoder generates the target sequence. An attention mechanism (Bahdanau et al., 2014) is used to generate the region of focus during decoding.

We employ NMT as the backbone of our paraphrasing model. In its simplest form our model exploits a one-to-one NMT architecture: the source English sentence is translated into k candidate foreign sentences and then back-translated into English. Inspired by multi-way machine translation which has shown performance gains over single-pair models (Zoph and Knight, 2016; Dong et al., 2015; Firat et al., 2016a), we also explore an alternative pivoting technique which uses multiple languages rather than a single one. Our model inherits advantages from NMT such as a small memory footprint and conceptually easy decoding (implemented as beam search). Beyond paraphrase generation, we experimentally show that the representations learned by our model are useful in semantic relatedness tasks.

Paraphrastic Embeddings The successful use of word embeddings in various NLP tasks has provided further impetus to use paraphrases. Wieting et al. (2015) take the paraphrases contained in PPDB and embed them into a low-dimensional space using a recursive neural network similar to Socher et al. (2013). In follow-up work (Wieting et al., 2016), they learn sentence embeddings based on supervision provided by PPDB. In our approach, embeddings are learned as part of the model and are available for any-length segments making use of no additional machinery beyond NMT itself.

3 Neural Paraphrasing

In this section we present PARANET, our **Para**phrasing model based on **Ne**ural Machine **T**ranslation. PARANET uses neural machine translation to first translate from English to a foreign pivot, which is then back-translated to English, producing a paraphrase. In the following, we briefly overview the basic encoder-decoder NMT framework and then discuss how it can be extended to paraphrasing.

3.1 NMT Background

In the neural encoder-decoder framework for MT (Sutskever et al., 2014; Bahdanau et al., 2014; Luong et al., 2015), the encoder, a recurrent neural network (RNN), is used to compress the meaning of the source sentence into a sequence of vectors. The decoder, a conditional RNN language model, generates a target sentence word-by-word. For the language pair, an encoder takes in a source sentence $X = \{x_1, ..., x_{T_X}\}$, as a sequence of linguistic symbols and produces a sequence of context vectors $C = \{h_1, ...h_{T_X}\}$. PARANET uses a bi-directional RNN, where each context vector h_t is the concatenation of the forward and the backward RNN's hidden states at time t.

The decoder is a conditional RNN language model that produces, given the source sentence, a probability distribution over the translation. At each time step t', the decoder's hidden state is updated:

$$z_{t'} = RNN(z_{t'-1}, y_{t'-1}, c_{t'}) \quad (1)$$

The update uses the previous hidden state $z_{t'-1}$, the previous target symbol $y_{t'-1}$ and the time dependent context $c_{t'}$, which is computed by an attention mechanism $\alpha_{t,t'}$ over the source sentences' context vectors:

$$c_{t'} = \sum_{t=1}^{T_x} \alpha_{t',t} h_t \quad (2)$$

$$\alpha_{t',t} \propto e^{f(z_{t'-1}, h_t)} \quad (3)$$

g is a feedforward neural network with a softmax activation function in the output layer which returns the probability of the next target symbol. The probability of the target sentence $Y = \{y_1, ..., y_{T_Y}\}$, is the product of the probabilities of the symbols within the sentence:

$$P(Y|X) = \prod_{t'=1}^{T_Y} P(y_{t'}|y_{<t'}, X) \quad (4)$$

3.2 Pivoting

Pivoting is often used in machine translation to overcome the shortage of parallel data, i,e., when there is not a translation path from the source language to the target. Instead, pivoting takes advantage of paths through an intermediate language. The idea dates back at least to Kay (1997), who observed that ambiguities in translating from one language onto another may be resolved if a translation into some third language is available, and has met with success in traditional phrase-based SMT (Wu and Wang, 2007; Utiyama and Isahara, 2007) and more recently in neural MT systems (Firat et al., 2016b).

In the case of paraphrasing, there is not a path from English to English. Instead, a path from English to French to English can be used. In other words, we translate a source sentence into a pivot language and then translate the pivot back into the source language. Pivoting using NMT ensures that the entire sentence is considered when choosing a pivot. The fact that contextual information is considered when translating, allows for a more accurate pivoted sentence. It also places greater emphasis on capturing the meaning of the sentence, which is a key part of paraphrasing.

A naive approach to pivoting is one-to-one back-translation. The source English sentence E_1, is translated into a single French sentence F. Next, F is translated back into English, giving a probability distribution over English sentences, E_2. This translation distribution acts as the paraphrase distribution $P(E_2|E_1, F)$:

$$P(E_2|E_1, F) = P(E_2|F) \quad (5)$$

One-to-one back-translating offers an easy way to paraphrase, because existing NMT systems can be used with no additional training or changes. However, there are several disadvantages; for example the French sentence F must fully capture the exact meaning of E_1, as E_1 and E_2 are conditionally independent given F. Since there is rarely a clear one-to-one mapping between sentences in different languages, information about the source sentence can be lost, leading to inaccuracies in the paraphrase probabilities. To avoid this, we propose back-translating through multiple sentences within one and multiple foreign languages.

Multi-pivoting PARANET pivots through the set of K-best translations $\mathcal{F} = \{F_1, ...F_K\}$ of E_1. This ensures that multiple aspects (semantic and syntactic) of the source sentence are captured. Moreover, multiple pivots provide resilience against a single bad translation, which would prevent one-to-one back-translation from producing accurate paraphrase probabilities.

Translating from multiple pivot sentences into one target sentence requires that the decoder be redefined. Firat et al. (2016b) propose several ways in which multiple pivot sentences can be incorporated into a NMT decoder. We extended their late averaging approach to incorporate weights. Consider the case of two pivot sentences from the same language, F_1 and F_2. Each translation path individ-

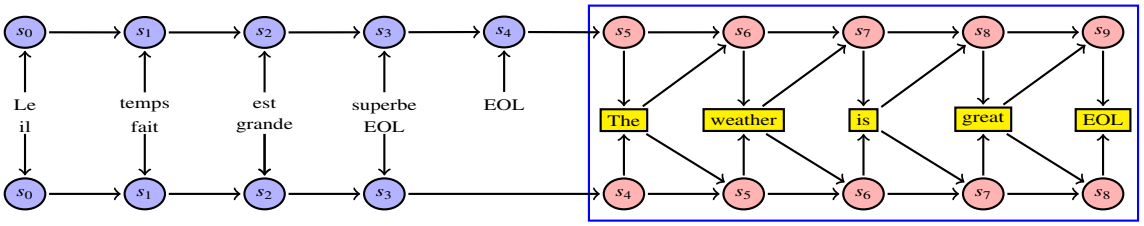

Figure 1: Late-weighted combination: two pivot sentences are simultaneously translated to one target sentence. Blue circles indicate the encoders, which individually encode the two source sentences. After the EOL token is seen, decoding starts (red circles). At each time step the two decoders produce a probability distribution over all words, which are then combined (in the yellow square) using Equation (6). From this combined distribution a word is chosen, which is then given as input to each decoder.

ually computes the distribution over the target vocabulary $P(y_{t'} = w|y_{<t'}, F_1)$ and $P(y_{t'} = w|y_{<t'}, F_2)$. Our *late-weighted combination* approach defines the path with respect to both translations as:

$$P(y_{t'} = w|y_{<t'}, F_1, F_2) = \lambda_1 P(y_{t'} = w|y_{<t'}, F_1) + \lambda_2 P(y_{t'} = w|y_{<t'}, F_2)$$

While Firat et al. (2016b) train a new model to capture these joint translations, we leave the model unchanged, instead treating PARANET as a *meta* encoder-decoder model (see Figure 1).

Unlike late averaging, PARANET assigns weights λ to each pivot sentence. These weights are set to the initial translation probabilities $P(F_i|E_1)$, thus capturing the model's confidence in the accuracy of the translation:

$$P(y_{t'}{=}w|y_{<t'}, F_1, F_2) = P(F_1|E_1)P(y_{t'}{=}w|y_{<t'}, F_1) + P(F_2|E_1)P(y_{t'}{=}w|y_{<t'}, F_2)$$

Which can be trivially extended to include all translations from the K-best list:

$$P(y_{t'} = w|y_{<t'}, \mathcal{F}) = \sum_{i=1}^{K} P(\mathcal{F}_i|E_1) \cdot P(y_{t'} = w|y_{<t'}, \mathcal{F}_i) \quad (6)$$

To ensure a probability distribution, we normalize the K-best list $\mathcal{F}$, such that the translation probabilities sum to one.

Multi-lingual Pivoting PARANET further expands on the multi pivot approach by pivoting not only over multiple sentences from one language, but also over multiple sentences from multiple languages. Multi-lingual pivoting has been recently shown to improve translation quality (Firat et al., 2016b), especially for low-resource language pairs. Here, we hypothesize that it will also lead to more accurate paraphrases.

Multi-lingual pivoting requires a small extension to late-weighted combination. We illustrate with German as a second language. First, the source sentence is translated into a K-best list of French $\mathcal{F}^{Fr}$, and a K-best list of German $\mathcal{F}^{De}$. Late-weighted combination is then applied, producing $P(y_{t'} = w|y_{<t'}, \mathcal{F}^{Fr})$ and $P(y_{t'} = w|y_{<t'}, \mathcal{F}^{De})$. These two output distributions are averaged, producing a multi-sentence, multilingual paraphrase probability:

$$P(y_{t'} = w|y_{<t'}, \mathcal{F}^{Fr}, \mathcal{F}^{De}) = \frac{1}{2}\left(P(y_{t'} = w|y_{<t'}, \mathcal{F}^{Fr}) + P(y_{t'} = w|y_{<t'}, \mathcal{F}^{De})\right)$$

which is used to obtain probability distributions over sentences:

$$P(E_2|E_1) = \prod_{t'=1}^{T_{E_2}} P(y_{t'}|y_{<t'}, \mathcal{F}^{Fr}, \mathcal{F}^{De}) \quad (7)$$

This can be trivially generalized to multiple languages. In this paper we use up to three.

3.3 PARANET Applications

The applications of PARANET are many and varied. We discuss some of these here and present detailed experimental evidence in Section 4. PARANET can be readily used for paraphrase detection (the task of analyzing two text segments and determining if they have the same meaning), by computing Equation (7). In addition, it can identify which linguistic units are considered paraphrases and to what extent. PARANET's explanatory power stems from the attention mechanism inherent in the NMT systems.

In encoder-decoder models, attention is used during each step of decoding to indicate which are the relevant source words. In our case, each word of the paraphrase attends to words within the pivot sentence and each word in the pivot sentence attends to words within the source sentence. By summing out the weighted pivot sentence, it is possible to see the attention from paraphrase to

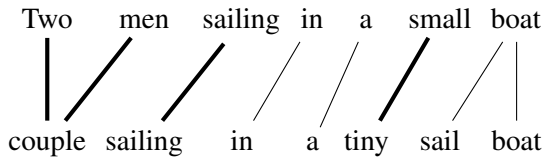

Figure 2: Attention between two sentences. Line thickness indicates the strength of the attention.

source:

$$\alpha(E_2^i, E_1^j, \mathcal{F}) = \sum_{F}^{\mathcal{F}} \left(P(E_2|E_1, F) \cdot \sum_{m}^{T_F} (\alpha_{i,m}^{E_2,F} \cdot \alpha_{m,j}^{F,E_1}) \right) \quad (8)$$

An example shown in Figure 2 where attention has successfully identified the semantically equivalent parts of two sentences. Beyond providing interpretable paraphrasing, attention scores can be used as features in both generation and classification tasks.

Furthermore, PARANET can be readily used to perform text generation (via the NMT decoder) without additional resources or parameter estimation. It also learns phrase and sentence embeddings for free without any model adjustments or recourse to resources like PPDB.

4 Experiments

We evaluated PARANET in several ways: (a) we examined whether the paraphrases learned by our model correlate with human judgments of paraphrase quality; (b) we assessed PARANET in paraphrase and similarity detection tasks; and (c) in a sentence-level paraphrase generation task. We first present details on how PARANET and comparison models were trained and then discuss our results.

4.1 Neural Machine Translation Training

We used Groundhog[1] as the implementation of the NMT system for all experiments. We generally followed the settings and training procedure from previous work (Bahdanau et al., 2014; Sennrich et al., 2016a). As such, all networks have a hidden layer size of 1000, and an embedding layer size of 620. During training, we used Adadelta (Zeiler, 2012), a minibatch size of 80, and the training set was reshuffled between epochs. We trained a network for approximately 7 days on a single GPU, then the embedding layer was fixed and training continued, as suggested in Jean et al. (2015a), for 12 hours. Additionally, the softmax was calculated over a filtered list of candidate translations. Following Jean et al. (2015a), we set the common vocabulary size as 10000 and 25 uni-gram translations, using a bilingual dictionary based on fast-align (Dyer et al., 2013).

In our experiments, we used up to six encoder-decoder NMT models (three pairs); English→French, French→English, English→Czech, Czech→English, English→German, German→English. All systems were trained on the available training data from the WMT15 shared translation task (4.2 million, 15.7 million, and 39 million sentence pairs for EN↔DE, EN↔CS, and EN↔FR, respectively). For EN↔DE and EN→CS, we also had access to back-translated monolingual training data (Sennrich et al., 2016a), which we also used in training. The data was pre-processed using standard pre-processing scripts found in MOSES (Koehn et al., 2007). Rare words were split into sub-word units, following Sennrich et al. (2016b). BLEU scores for each NMT system can be seen in Table 1.

4.2 Statistical Machine Translation Training

Throughout our experiments we compare PARANET against a paraphrase model trained with a commonly used Statistical Machine Translation system (SMT), which we henceforth refer to as PARASTAT. Specifically, for each language pair used, an equivalent IBM Model 4 phrase-based translation model was trained. Additionally, an Operation Sequence Model (OSM) was included, which has been shown to improve the performance of SMT systems (Durrani et al., 2011). SMT translation models were implemented using both GIZA++ (Och and Ney, 2003) and MOSES (Koehn et al., 2007) and were trained using the same pre-processed bilingual data provided to the NMT systems. The SMT systems used a KenLM 5-gram language model (Heafield, 2011), trained on the mono-lingual data from WMT 2015. For all languages pairs, both KenLM and MOSES were trained using the standard settings. BLEU scores for the SMT systems are given in Table 1.

Under the SMT models, paraphrase probabilities were calculated analogously to Equation (7):

$$P(E_2|E_1, \mathcal{F}) = \sum_{F}^{\mathcal{F}} P(E_2|F) P(F|E_1) \quad (9)$$

where $P(E_2|F)$ and $(F|E_1)$, are defined by the phrase based translation model, and $\mathcal{F}$ denotes the K-best translations of E_1, whose probabilities are normalized. Unlike PARANET these pivot sentences have to be combined outside of the decoder.

[1] github.com/sebastien-j/LV_groundhog

Direction	F→E		E→F	
System	SMT	NMT	SMT	NMT
French	0.241	0.201	0.233	0.271
German	0.207	0.282	0.208	0.248
Czech	0.216	0.197	0.145	0.176

Table 1: BLEU scores (WMT 2015 test set) for SMT and NMT models (foreign to English (F→E) and English to foreign (E→F) directions).

4.3 Correlation with Human Judgments

The PPDB 2.0 Human Evaluation data set is a sample of paraphrase pairs taken from PPDB which have been human annotated for semantic similarity (Pavlick et al., 2015). 26,455 samples were taken from range of syntactic categories, resulting in paraphrase candidates varying from single words to multi-word expressions. Each paraphrase pair was judged by five people on a 5-point scale. Ratings were then averaged giving each paraphrase pair a score between 1 and 5.

Using this dataset we measure the correlation (Spearman ρ) between (length normalized) PARANET probabilities (Equation (7)) assigned to paraphrase pairs and human judgments. Figure 3 shows correlation coefficients for all language pairs using a single foreign pivot and 200 pivots. Across all language combinations multiple pivots[2] achieve better correlations, with the German, Czech pair performing best with $\rho = 0.53$. For comparison, Pavlick et al. (2015) report a correlation of $\rho = 0.41$ using Equation (9) and PPDB (Ganitkevitch et al., 2013). The latter contains over 100 million paraphrases and was constructed over several English-to-foreign parallel corpora including Europarl v7 (Koehn, 2005) which contains bitexts for the 19 European languages.

Following Pavlick et al. (2015), we next developed a supervised scoring model. Specifically, we fit a decision tree regressor on the PPDB 2.0 dataset using the implementation provided in scikit-learn (Pedregosa et al., 2011). To improve accuracy and control over-fitting we built an ensemble of regression trees using the Extra-Trees algorithm (Geurts et al., 2006) which fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset. In our experiments 1,000 trees were trained to minimize mean square error. The regressor was trained with the following basic features: sentence length,

[2]Across tasks and datasets we find that multiple pivots outperform single pivots. We omit these comparisons from subsequent experiments for the sake of brevity.

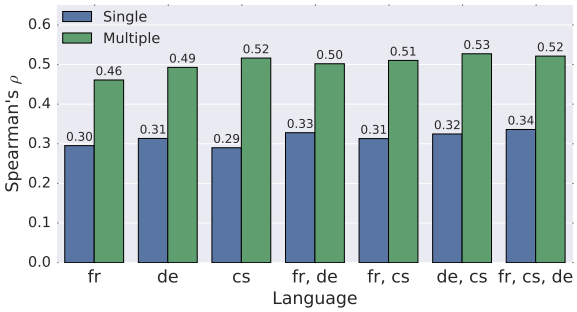

Figure 3: Correlation of PARANET predictions against human ratings for paraphrase pairs. Comparison using single and multiple pivots, across language combinations.

1-4 gram string similarity, the paraphrase probability $P(E_2|E_1)$, the language model score $P(E_1)$, cosine distance of the sentence vectors, as calculated by the encoder. To address the problem of rare sentences receiving low probabilities regardless of the source sentence, we create an inverse weighting by $P(E_2|E_2)$, which approximates how difficult it is to recover E_2:

$$pscore(E_2, E_1) = \frac{P(E_2|E_1)}{P(E_2|E_1) + P(E_2|E_2)} \quad (10)$$

Two features reflect the alignment between candidate paraphrases. We built an alignment matrix according to Equation (8), and used the mean of the diagonal as feature. This acts as a proxy of how much movement there is between two paraphrases. The second feature is the number of unaligned words which we compute by calculating hard alignments between the two paraphrases.

Regressors varied with respect to how $P(E_2|E_1)$ was computed, keeping the string based features the same. Equations (7) and (9) were used to calculate paraphrase probability for PARANET and PARASTAT, respectively. For both models beam search (with width set to 100) was used to generate the K-best list. For each language, the K-best list is the union of the 100-best list of E_1 and the 100-best list of E_2, giving a maximum of 200 pivot sentences. As set out in Pavlick et al. (2015) evaluation is done using cross validation: in each fold, we hold out 200 phrases. Table 2 presents results for PARANET and PARASTAT using different languages as pivots. PARANET outperforms PARASTAT across the board. Furthermore, despite using fewer features and pivot languages, it obtains a closer correspondence to human data compared to PPDB 2.0 (Pavlick et al., 2015).

Model	PARASTAT	PARANET
fr	0.574	0.700
de	0.638	0.710
cz	0.564	0.713
de,fr	0.566	0.722
de,cz	0.640	0.731
fr,cz	0.569	0.724
fr, cz, de	0.633	**0.735**
PPDB 2.0	0.713	

Table 2: Correlation (Spearman ρ) of supervised models against human ratings for paraphrase pairs. Boldface indicates the best performing model.

4.4 Paraphrase Identification and Similarity

The SemEval-2015 shared task on Paraphrase and Semantic Similarity In Twitter (PIT) uses a training and development set of 17,790 sentence pairs and a test set of 972 sentence pairs. By design, the dataset contains colloquial sentences representing informal language usage and sentence pairs which are lexically similar but semantically dissimilar. Sentence pairs were crawled from Twitter's trending topics and associated tweets (see Xu et al. (2014) for details). The shared task consists of a (binary) paraphrase identification subtask (i.e., determine whether two sentences are paraphrases) and an optional semantic similarity task (i.e., determine the similarity between two sentences on a scale of 1–5, where 5 means completely equivalent and 1 not equivalent).

We trained a decision tree regressor on the PIT-2015 similarity dataset using the features described above. Once trained, the decision tree regressor can be readily applied to the semantic similarity subtask. For the paraphrase detection subtask, we use the same model and apply a threshold (optimized on the validation set) such that those pairs that are over this threshold are deemed paraphrases.

Tables 3 and 4 present our results on the two subtasks together with previously published results. We evaluate system performance on the detection task using F1 (the harmonic mean of precision and recall). For semantic similarity, system outputs are compared by Pearson correlation against human scores. The first block in the tables summarize results for PARANET and PARASTAT using different languages as pivots. The second block includes three baselines provided by the organizers of the shared task: a random baseline, a logistic regression baseline with minimal

Model	PARASTAT	PARANET
fr	0.613	0.624
de	0.616	0.620
cz	0.620	0.622
de, fr	0.602	0.622
de, cz	0.606	0.615
fr, cz	0.600	**0.634**
fr, cz, de	0.596	0.620
random	0.266	
WTMF	0.536	
logistic reg	0.589	
ASOBEK	0.674	
MITRE	0.667	

Table 3: Paraphrase detection results (F1) on the PIT-2015 data set. Boldface indicates the best performing paraphrasing model.

Model	PARASTAT	PARANET
fr	0.540	0.569
de	0.543	**0.571**
cz	0.547	0.569
de, fr	0.543	0.569
de, cz	0.540	0.570
fr, cz	0.546	0.568
fr, cz, de	0.539	0.568
random	0.017	
WTMF	0.350	
logistic reg	0.511	
ASOBEK	0.475	
MITRE	0.619	

Table 4: Semantic similarity results (Pearson) on the PIT-2015 data set. Boldface indicates the best performing paraphrasing model.

n-gram word overlap features; and a model which uses weighted matrix factorization (WTMF) and has access to dictionary definitions provided in WordNet, OntoNotes, and Wiktionary (Guo and Diab, 2012). The last two rows show the highest scoring systems: ASOBEK (Eyecioglu and Keller, 2015) ranked 1st in the identification subtask and MITRE (Zarrella et al., 2015) in the similarity subtask. Whereas ASOBEK uses knowledge-lean features based on word and character n-gram overlap, MITRE is a combination of multiple systems including mixtures of string matching metrics, alignments using tweet-specific word representations, and recurrent neural networks.

As can be seen, PARANET achieves better similarity and detection score than all baselines and PARASTAT, for any combinations of lan-

Model	PARASTAT	PARANET
fr	0.657	0.682
de	0.666	0.678
cz	0.649	0.688
de, fr	0.665	0.684
de, cz	0.662	0.687
fr, cz	0.654	**0.690**
fr, cz, de	0.658	0.689
Tokencos	0.587	
DLS@CU	0.801	

Table 5: Results on the Semeval-2015 semantic similarity dataset. Boldface indicates the best performing paraphrasing model.

guages. This is particularly impressive as the translation models were trained on very dissimilar data. Compared to the state of the art, PARANET fares worse, however our model was not particularly optimized on the PIT-2015 dataset which was merely used as a testbed for a fair comparison. It is thus reasonable to assume that taking into account more elaborate features (e.g., based on character embeddings) would improve performance. The highest semantic similarity score is obtained with PARANET trained using German data. The highest scoring paraphrase detection model was PARANET trained on French and Czech data. Interestingly, using multiple pivot languages seems to offer small improvements in most cases. The languages selected as pivots in our experiments were somewhat ad-hoc. We expect to get more mileage if these are selected from the same language family or with more linguistic insight (e.g., morphologically rich vs. poor).

4.5 Semantic Textual Similarity

In semantic textual similarity (STS), systems rate the degree of semantic equivalence between two text snippets. We present results on the Semeval-2015 English subtask which contains sentences from a wide range of domains, including newswire headlines, image descriptions, and answers from Q&A websites. The training/test sets consist of 11,250 and 3,000 sentence pairs, respectively. Sentence pairs are rated on a 1–5 scale, with 5 indicating they are completely equivalent.

We used the decision tree regressor with the same features described in the previous section. Again, we experimented with one, two, and three languages as pivots, and compared PARANET and PARASTAT directly. Our results are summarized in Table 5. The third block in the table presents a simple cosine-based baseline provided by the organizers (Tokencos) and the top-performing system (DLS@CU) which uses PPDB paraphrases to identify semantically similar words and word2vec embeddings trained on approximately 2.8 billion tokens (Sultan et al., 2014).

PARANET outperforms PARASTAT on all languages and language combinations. Both systems outperform the Semeval baseline but are worse compared to the top scoring system. We see for PARANET Czech achieves the highest scores, this could be in part due to Czech non-strict word order, which allows paraphrases that are simple rearrangements not be penalized.

4.6 Paraphrase Generation

Finally, we evaluated PARANET (and PARASTAT) in a paraphrase generation task. We created sentential paraphrases for three (parallel monolingual) datasets representative of different domains and genres: (a) the Multiple-Translation Chinese (MTC) corpus (Huang et al., 2002) contains news stories from three sources of journalistic Mandarin Chinese text translated into English by 4 translation agencies; we sampled 1,000 sentences for training and testing, respectively (each source sentence had an average of 4 paraphrases); (b) the Jules Vernes Twenty Thousand Leagues Under the Sea novel (Leagues) corpus (Pang et al., 2003) contains two English translations of the French novel; we sampled 500 sentences for training/testing (each source sentence had one paraphrase); and (c) the Wikianswers corpus (Fader et al., 2013) which contains questions taken from the website[3] wiki answers; we sampled 1,000 questions for training/testing (each question has on average 21 paraphrases).

In order to select the best paraphrase candidate for a given input sentence, PARASTAT was optimized on the training set using Minimum Error Training (MERT, Och and Ney (2003)). MERT integrates automatic evaluation metrics such as BLEU into the training process to achieve optimal end-to-end performance. Naively optimizing for BLEU, however, will result in a trivial paraphrasing system heavily biased towards producing identity "paraphrases". Sun and Zhou (2012) introduce iBLEU which we also adopt. iBLEU penalizes paraphrases which are similar to the source

[3] http://wiki.answers.com/

Model	PARASTAT	PARANET
fr	0.280	0.299
de	0.282	0.295
cz	0.280	0.291
Gold	0.599	

Table 6: Mean iBLEU across three datasets.

sentence and rewards those close to the target:

$$iBLEU(s, r_s, c) = \alpha BLEU(c, r_s) - (1-\alpha) BLEU(c, s)$$

where s, is the source sentence, r_s, is the target and c is the candidate paraphrase. $(1-\alpha)BLEU(c,s)$, measures the originality of the candidate paraphrase, $BLEU(c, r_s)$ measures semantic adequacy, and α is a tuning parameter which balances the two. Sentence level BLEU is calculated using plus one smoothing (Lin and Och, 2004).

PARANET relies on a relatively simple architecture which is trained end-to-end with the objective of maximizing the likelihood of the training data. Since evaluation metrics cannot be straightforwardly integrated into this training procedure, we reranked the k-best paraphrases obtained from PARANET using a simple classifier which favors sentences which are dissimilar to the source. Specifically, we trained a decision tree regression model with iBLEU as the target variable using the same features described in Section 4.4. Examples of paraphrases generated by PARANET are shown in the Appendix.

System output was assessed automatically using iBLEU with human-written paraphrases as reference. In addition, we evaluated the generated text by eliciting human judgments via Amazon Mechanical Turk. We randomly selected 100 source sentences from each data set and generated output with PARANET and PARASTAT (using German as a pivot). We also included a randomly selected human paraphrase as a goldstandard. Workers (self-reported native English speakers) were asked to rank the three paraphrases from best to worst (ties were allowed) in order of semantic equivalence (does the paraphrase convey the same meaning as the source?) and fluency (is the description written in well-formed English?). Participants were explicitly told to give high ranks to output demonstrating a fair amount of paraphrasing and low ranks to trivial paraphrases (e.g., deletion of articles or punctuation). We collected 5 responses per input sentence.

Table 6 summarizes our results across the three datasets. For the sake of brevity, we only show results with one pivot language since combinations performed slightly worse for both models. We set $\alpha = 0.8$ for iBLEU as we experimentally found it offers the best trade-off between semantic equivalence and dissimilarity. As an upper-bound we also measure iBLEU amongst the gold paraphrases provided by humans. Again, we observe that PARANET has a slight advantage over PARASTAT in terms of iBLEU, however both systems tend to paraphrase less compared to the gold-standard. Table 7 shows the mean ranks given to these systems by human subjects. An Analysis of Variance (ANOVA) revealed a reliable effect of system type. Post-hoc Tukey tests showed that PARANET is significantly ($p < 0.01$) better than PARASTAT across datasets; PARANET is also significantly ($p < 0.01$) better than the the gold standard on both MTC and the Wikianswers dataset. We attribute this to the noisy nature of these two datasets which contain a wealth of paraphrases, a few of which are ungrammatical, contain typos or abbreviations leading to low scores among humans.

Model	Wikianswers	Leagues	MTC	All
PARASTAT	2.09	2.38	2.23	2.26
PARANET	**1.86**	1.94	**1.70**	**1.83**
Humans	2.17	**1.81**	2.0	2.0

Table 7: Mean Rankings given to paraphrases by human participants (a lower score is better).

5 Conclusions

In this work we presented PARANET, a neural paraphrasing model based on bilingual pivoting. Experimental results across several tasks (similarity prediction, paraphrase identification, and paraphrase generation) show that PARANET outperforms conventional paraphrasing methods. In the future, we plan to exploit the attention scores more directly for extracting paraphrase pairs (in analogy to PPDB) and as features for classification tasks (e.g., textual entailment). We would also like to investigate how PARANET can be adapted using reinforcement learning (Ranzato et al., 2016) to text generation tasks such as simplification and sentence compression.

Acknowledgments The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1; Mallinson) and the European Research Council (award number 681760; Lapata).

Appendix

Tables 8–10 show examples of PARANET output on the Wikianswers, Leagues, and MTC datasets.

Wikianswers
a. How many calories in a handful of strawberries? b. The number of calories in a handful of strawberries.
a. Beauty is not in the eye of the beholder. b. Beauty is not in the mind of the viewer.
a. What is the importance of employee satisfaction in an organization? b. What is the significance of staff satisfaction at an organisation?
a. What is the difference between electrical power and electrical energy? b. What is the difference between electrical energy and electrical power?
a. How many high tides happen at a given coast in any 24 hour period? b. How many high tides occur on a certain coast in 24 hours?
a. What is a beverage that starts with the letter p? b. What is a drink that begins with the letter p?
a. What Swiss mathematician and teacher was responsible for instituting the use of the symbol for π in mathematical notation? b. What Swiss mathematicians and teachers were responsible for the introduction of the symbol for π in math notation?
a. How do you make a pina colada? b. How do you do a Pina colada?
a. What is the difference between a captain and a skipper? b. What is the difference between being a captain and skipper?

Table 8: Sentences marked (a) are the input and (b) are PARANET paraphrases.

Leagues
a. "Faith i should never have believed it," said Conseil. b. "Faith, I never would have believed", Conseil said.
a. "I owed myself this revenge!" Said the Captain to the Canadian. b. "I am indebted to this revenge!" the captain told the Canadian.
a. "Well, sir, you will only get your deserts." b. "Well, sir, you are only getting your deserts."
a. "That's what I've been telling you Ned." b. "That's what I said, Ned."
a. Very much embarrassed, after having vainly exhausted our speaking resources, I knew not what part to take, when Conseil said: "if master will permit me I will relate it in German." b. It was very embarrassing that I had used up our speaking time, and I did not know what to do, as Conseil said: "If the Masters allow me, I shall refer to German."
a. Almost every day the panels in the lounge were open for some hours, and our eyes never tired of probing the mysteries of the underwater world. b. Almost every day, the panels opened in the lounge for a few hours, and our eyes never tired, the secrets of the underwater world.
a. I bowed, and desired to be conducted to the cabin destined for me. b. I bow to and wish I headed to the cabin for me.
a. I had one question left to address to Captain Nemo. b. I had a question left to Captain Nemo.
a. "I have not the foggiest notion, Professor Aronnax." b. I have no idea, Professor Aronnax.

Table 9: Sentences marked (a) are the input and (b) are PARANET paraphrases.

MTC
a. China expresses strong dissatisfaction over the Japanese leader's move this time. b. China expresses a strong dissatisfaction over Japanese leader's move.
a. We will accelerate the drafting of telecommunications legalization, amend the law of post and the regulations governing wireless telecommunications. b. We will speed up the design of telecommunications, change the law and regulations governing wireless telecommunication.
a. Liu said: the poverty-stricken areas are badly hit in the first stage of this year's floods and many counties and cities are listed as the poorest ones in the country. b. Liu said: poverty-stricken areas are hit hard in the first phase of this year's flooding and many counties and towns are listed as the poorest in the country.
a. (London, AP) The British government is working on resolving the increasingly serious problems of street crimes and will strengthen patrolling police. b. London, AP The British government is working to resolve the increasingly serious problems of street crime and will strengthen patrols.
a. Kida said that the dead killed by the heat wave were mostly old people with heart diseases. b. Kida said the dead by heatwave were mostly old people with heart disease.
a. Growth of Mobile Phone users in Mainland China to Slow Down. b. Growth of Mobile Phone users in Mainland China on Slow Down.
a. The survey report said that in the first six months of last year 18 sandstorms struck Beijing and they all originated from Inner Mongolia where 60% of the land is desert. b. The report said that 18 sandstorms struck Beijing in the first six months of last year, and they were all from Inner Mongolia, where 60% of the desert is desert.
a. The World Cup co-host by Japan and South Korea, will inaugurate on May 31. b. The World Cup, co-host Japan and South Korea, will open on May 31.
a. Two days ago, President Bush seemed opposed to this idea when he held talks with Sharon. b. Two days ago President Bush opposed this idea when he talks to Sharon.
a. Russia Faces Population Crisis. b. Russia's demographics problem.
a. Computer Crimes Cost US billions of Dollars Last Year. b. Computer Crimes Cost American Billions of Dollars.
a. However, many sports associations in Chile hope to cooperate with China not just for the table tennis alone. b. However, many sports federations in Chile are hoping to collaborate with China, not only for the table tennis players.

Table 10: Sentences marked (a) are the input and (b) are PARANET paraphrases.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Associa-*

tion for Computational Linguistics, pages 597–604, Ann Arbor, Michigan.

Regina Barzilay. 2003. *Information Fusion for Multi-Document Summarization: Praphrasing and Generation*. Ph.D. thesis, Columbia University.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Baltimore, Maryland.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, New York City, USA.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Honolulu, Hawaii.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1045–1054, Portland, Oregon, USA.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia.

Asli Eyecioglu and Bill Keller. 2015. Twitter paraphrase identification with simple overlap features and svms. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 64–69, Denver, Colorado.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. *CoRR*, abs/1606.04164.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland.

Juri Ganitkevitch, Chris Callison-Burch, Courtney Napoles, and Benjamin Van Durme. 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1168–1179, Edinburgh, Scotland, UK.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Weiwei Guo and Mona Diab. 2012. A simple unsupervised latent semantics based approach for sentence similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 586–590.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Shudong Huang, David Graff, George Doddington, Linguistic Data Consortium, et al. 2002. *Multiple-translation Chinese corpus*. Linguistic Data Consortium, University of Pennsylvania.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015a. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015b. Montreal neural machine translation systems for wmt15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(1-2):3–23.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54, Sapporo, Japan.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Philip Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics, Main Volume*, pages 605–612, Barcelona, Spain.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):342–360.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127, Prague, Czech Republic.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 102–109, Edmonton, Canada.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremb. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 464–471, Prague, Czech Republic.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. DLS@CU: Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 241–246.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual smt system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112. Curran Associates, Inc.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association of Computational Linguistics – Volume 3, Issue 1*, pages 345–358.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *Proceedings of 4th International Conference on Learning Representations*, San Juan, Puerto Rico.

Kristian Woodsend and Mirella Lapata. 2014. Text rewriting improves semantic role labeling. *Journal of Artificial Intelligence Research*, 51:133–164.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics*, 2:435–448.

Guido Zarrella, John Henderson, Elizabeth M. Merkhofer, and Laura Strickhart. 2015. Mitre: Seven systems for semantic similarity in tweets. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 12–17, Denver, Colorado.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008. Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of ACL-08: HLT*, pages 780–788, Columbus, Ohio.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California.